University of North Carolina at Chapel Hill
The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series

Year 2012 *Paper* 27

A Multistage Non?inferiority Study Analysis Plan to Evaluate Successively More Stringent Criteria for a Clinical Trial with Rare Events

Siying Li* Gary G. Koch†
Todd A. Schwartz DrPH‡

*University of North Carolina at Chapel Hill, sili@bios.unc.edu
†The University of North Carolina at Chapel Hill, bcl@bios.unc.edu
‡University of North Carolina at Chapel Hill, tschwart@email.unc.edu

A Multistage Non?inferiority Study Analysis Plan to Evaluate Successively More Stringent Criteria for a Clinical Trial with Rare Events

Siying Li, Gary G. Koch, and Todd A. Schwartz DrPH

Abstract

We address a multistage clinical trial to assess a sequence of hypotheses in the noninferiority and also rare events setting. Three successive hypotheses are used to evaluate whether the new treatment meets the criteria for new drug approval. Sample sizes for a five stage trial for all hypotheses are calculated using Poisson and Logrank sample size methods. Three strategies and corresponding analysis plans are developed to evaluate the sequential hypotheses. Simulations show the design is satisfactory with respect to controlled Type I error, sufficient power, and early success at interim analyses.

A Multistage Non-inferiority Study Analysis Plan to Evaluate Successively More Stringent Criteria for a Clinical Trial with Rare Events

Ms. Siying Li, Dr. Gary Koch, Dr. Todd Schwartz

Abstract

We address a multistage clinical trial to assess a sequence of hypotheses in the non-inferiority and also rare events setting. Three successive hypotheses are used to evaluate whether the new treatment meets the criteria for new drug approval. Sample sizes for a five stage trial for all hypotheses are calculated using Poisson and Logrank sample size methods. Three strategies and corresponding analysis plans are developed to evaluate the sequential hypotheses. Simulations show the design is satisfactory with respect to controlled Type I error, sufficient power, and early success at interim analyses.

Key words: Non-inferiority; interim analysis; adaptive design; Poisson sample size method; multiple comparisons.

1. Introduction

This paper discusses a multi-stage analysis plan for a randomized clinical trial to evaluate the cardiovascular risk in new therapies for Type 2 diabetes (FDA, 2008).New treatments for Type 2 diabetes can provide a reduced risk of micro-vascular complications, but possibly elevated risk of cardiovascular disease. The trial design has a multistage non-inferiority objective to assess the cardiovascular risk of the new drug; i.e., given the benefits of the new treatment, if the cardiovascular risk is less than a pre-specified margin (often larger than 1), then

the new treatment would have favorable benefit to risk profile (Rothmann, Li, Chen, Chi, Temple, & Tsou, 2003).

The endpoint of interest for the trial is major adverse cardiac events (MACE), defined as mortality, myocardial infarction, or stroke. In the patient population using type 2 diabetes drugs, MACE is a rare event. To enable the trial to be cost effective in the setting of rare events, we evaluate a sequence of hypotheses instead of a single hypothesis (Denne & Koch, 2001). For the first step, we evaluate whether the relative risk is less than 1.8; if successful, we evaluate whether the relative risk is less than 1.3 as the primary second step; if successful, we optionally evaluate the superiority hypothesis, that the relative risk is less than 1.0. If we experience failure at any step, then we stop the trial and make a conclusion based on the data observed through the step of the sequence where the process was halted.

Furthermore, the multistage trial is designed to assess the hypotheses at interim analyses for early success. We also consider early stopping for futility if the conditional power at a specified stage indicates unacceptably low power for success at the final stage.

The numbers of events for sufficient power to address each hypothesis of the study are calculated using a Poisson sample size method, and simulations are used to evaluate the study design in terms of its Type I error and power.

2. Sample Size Method

The Poisson sample size method assumes that events come from independent Poisson distributions, as could occur for rare events. The sample size method for the Logrank test assumes events come from independent time to event distributions with constant hazard ratio. The Logrank method is a robust sample size method without assumptions for time to event distributions, whereas the Poisson method is mainly for exponential time to event distributions.

For the assessment of the risk ratio, Poisson events can be generated from a binomial distribution conditional on the total number of events, and this simplifies the simulation process. We will compare the sample size results from the Poisson method to the Logrank method for each hypothesis in order to justify the use of the Poisson method for a simpler generation of the trials with respect to their objectives.

2.1 Poisson Method

We assume that time to event in the treatment group follows an exponential distribution $Exp(\lambda_T)$ and time to event in the reference group follows an exponential distribution $Exp(\lambda_R)$. For an observed total of *y* events during total person-time exposure *T*, the exponential likelihood $\lambda^y e^{-\lambda T}$ is proportional to the Poisson likelihood $\frac{(\lambda T)^y}{y!}e^{-\lambda T}$, with Poisson parameter $\mu = \lambda T$.

We now assume that the number of events in treatment group y_T has a Poisson distribution $Poisson(\mu_T = \lambda_T T)$ and the number of events in reference group y_R has a Poisson distribution $Poisson(\mu_R = \lambda_R T)$.

The non-inferiority hypotheses in our study are stated in terms of the hazard ratio $\frac{\lambda_T}{\lambda_R}$ and can be written as

$$H_0: R = \frac{\lambda_T}{\lambda_R} > R_0, vs, H_A: R = \frac{\lambda_T}{\lambda_R} \leq R_A.$$

To address the hazard ratio $\frac{\mu_T}{\mu_R} = \frac{\lambda_T}{\lambda_R}$, for the situation where the treatment group and the reference group have the same total person-time of exposure, the Poisson method can provide the needed number of events for the specified significance level and power.

Given the total number of events $\mathrm{d} = y_T + y_R$, y_T has a binomial distribution, $Binomial(d, \frac{\frac{\lambda_T}{\lambda_R}}{\frac{\lambda_T}{\lambda_R}+1})$. Let $R = \frac{\lambda_T}{\lambda_R}$. So the test is equivalent to

$$H_0: \pi = \frac{R_0}{R_0+1} > \pi_0, vs, H_A: \pi = \frac{R_A}{R_A+1} \leq \pi_A,$$

where $\pi_0 = \left(\frac{1.8}{2.8}\right), \left(\frac{1.3}{2.3}\right)$ for $R_0 = 1.8, 1.3$, and $\pi_A = 0.5$ for $R_A = 1$.

Denoting $d' = \frac{[Z_\alpha\sqrt{\pi_0(1-\pi_0)}+Z_\beta\sqrt{\pi_A(1-\pi_A)}]^2}{(\pi_A-\pi_0)^2}$, the Poisson sample size is given by $d_{Poisson} = \frac{d'}{4}[1+\sqrt{\frac{2}{d'|\pi_A-\pi_0|}}]^2$, where Z_α is the $(1-\alpha)^{\text{th}}$ percentile of the standard normal distribution (Fleiss, 1981).

2.2 Logrank Method

For the Logrank method, we assume the counts of events in the treatment group follow a survival distribution, S_T, and the counts of events in the reference group follow a survival distribution, S_R. The cumulative hazard for the treatment group is $-\ln S_T$; similarly, the cumulative hazard for the reference group is $-\ln S_R$. The hazard ratio for comparing the two groups can be expressed as $\mathrm{R} = \frac{\ln S_T}{\ln S_R}$, which is constant over time. Denoting $R_S = \frac{R_0}{R_A}$, we have the Logrank sample size formula as $d_{Logrank} = \frac{(Z_\alpha+Z_\beta)^2(1+R_S)^2}{(1-R_S)^2}$, which is adapted from the Freedman method (Freedman, 1982) (Dann & Koch, 2004).

2.3 Sequence of Non-inferiority Hypotheses

In our study for assessing cardiovascular risk of anti-diabetic therapies, we have two important boundaries as our hypotheses margins according to the FDA guidance (FDA, 2008). If the upper one-sided 0.975 confidence limit for the hazard ratio comparing the treatment group to the reference group cannot exclude values exceeding 1.8, the new therapy is not worthy for approval; if the upper one-sided 0.975 confidence limit for the hazard ratio is less than 1.8 but larger than 1.3, the new therapy may be approved, but post-marketing trials will be required to

show that the hazard ratio is definitely less than 1.3 by having the corresponding one-sided 0.975 confidence limit below it (Parks, 2009).

Corresponding to the FDA requirements, we first demonstrate the event rate ratio is less than 1.8, the non-inferiority margin, by assessing the following hypothesis, $H_{01}: R = \frac{\lambda_T}{\lambda_R} > 1.8, vs, H_{A1}: R = \frac{\lambda_T}{\lambda_R} \leq 1.0.$

If successful, we then demonstrate the event rate ratio is less than 1.3, the second non-inferiority margin, as the primary step, which is, $H_{02}: R = \frac{\lambda_T}{\lambda_R} > 1.3, vs, H_{A2}: R = \frac{\lambda_T}{\lambda_R} \leq 1.0.$

As our study is a non-inferiority study, we optionally demonstrate the event rate ratio is less than 1.0, which is a superiority assessment, $H_{03}: R = \frac{\lambda_T}{\lambda_R} > 1.0, vs, H_{A3}: R = \frac{\lambda_T}{\lambda_R} \leq 0.8.$

2.4 Sample Size Comparisons

Table 1 Sample Size Comparisons for Poisson versus Logrank Methods, using one-sided α=.025 and 1-β=.9

H_0	H_A	Poisson	Logrank
>1.8	≤1.0	130	129
>1.3	≤1.0	627	618
>1.0	≤0.8	865	851

For the sequence of hypotheses stated in Section 2.3, we use both the Poisson and Logrank sample size formulas to calculate the sample size needed for one-sided Type I error of 0.025 and power of 0.9. For H_{01}, where the hazard ratio larger than 1.8 is tested versus the hazard ratio less than 1.0, the sample sizes are 130 and 129 via the Poisson and Logrank methods, respectively. For H_{02}, testing the hazard ratio larger than 1.3 versus the hazard ratio less than 1.0, the respective sample sizes are 627 and 618. For the superiority hypothesis H_{03}, the sample sizes are, respectively, 865 and 851. Thus, we can conclude that the Poisson and Logrank methods yield similar results for each hypothesis, which supports the use of the Poisson method in our simulation plan. Detailed sample size information for more alternative hypotheses is provided in Table A in the Appendix.

3. Three Strategies to Address Sequential Hypotheses and Analysis Plan

3.1 Strategies to assess a sequence of hypotheses and overall analysis plan

Three strategies are developed to address the sequence of hypotheses. The first is to design a single study to assess all 3 hypotheses sequentially. The second strategy is to add interim analyses for success for each hypothesis so that we can achieve early success through rejection of the null hypotheses at the early interim analyses. The third strategy is to allow possible early stopping for futility with respect to unacceptably low conditional power to succeed at the final stage for each hypothesis. The last two strategies can shorten the length of the trial for its conclusions since the study outcome of our study is a rare event.

Table 2 Analysis Plan for a Study with 900 Events

Stage	n_i	N_i	H_{01}:R>1.8	H_{02}:R>1.3	H_{03}:R>1.0
1	100	100	Interim	Interim	Interim
2	100	200	Final	Interim	Interim
3	250	450	N/A	Interim	Interim
4	250	700	N/A	Final	Interim
5	200	900	N/A	N/A	Final

n_i: Number of new events at Stage *i*;
N_i: Cumulative number of events by Stage *i*.

According to the Poisson sample size we calculated earlier, the required sample sizes for H_{01}, H_{02}, H_{03} respectively are 130, 627 and 865 events. We thus choose the total sample size for the study as 900 events. The trial is divided into a five-stage trial, and the interim analyses are at 1/9, 2/9, 1/2, 7/9, and 1/1 of the total sample size, which corresponds to 100, 200, 450, 700 and 900 events. The final analysis for H_{01} is at Stage 2 (200 events) so as to exceed 0.90 power for that hypothesis since the entire study fails if there is failure to reject H_{01}; the final analysis for H_{02} is at Stage 4 (700 events); and the final analysis for H_{03} is at Stage 5 (900 events). The overall analysis plan is illustrated in Table 2.

Possible early stopping for futility is assessed at Stage 3 for H_{02} and H_{03}, and also at Stage 4 for H_{03}. The futility rule for early stopping will be discussed later in Section 4.2.

The analysis plan below will be based on the third strategy, which also enables the power simulation of the other two strategies.

3.2 Analysis Plan for H_{01}: R>1.8

For $H_{01}: R_0 > 1.8$, we assess the power of rejecting the null hypothesis when the true hazard ratio under the alternative H_{A1} is each of $R_A = 1.3, 1.15, 1.0, 0.9, \text{and } 0.8$.

At Stage 1 of 100 events, we assess H_{01} at one-sided $\alpha_{11} = 0.0125$.

At Stage 2, given no rejection at Stage 1, then at 200 events, we assess H_{01} at one-sided $\alpha_{21} = 0.0150$ with a strategy like the Hochberg method with $\alpha_1 = 0.025$ (Hochberg, 1988). With this Hochberg method, we reject H_{01} if either the maximum of p-values from Stage 1 and Stage 2 is less than α_1, or the p-value at Stage 2 is less than α_{21}, and if so, we conclude the hazard ratio is less than 1.8.

The cumulative power is computed for rejection of H_{01} by Stage 1 and 2, respectively, with that by Stage 2 being the type I error when H_{01} holds.

3.3 Analysis Plan for H_{02}: R>1.3

The rejection of H_{02} at any stage requires rejection of H_{01} at that stage or a previous stage.

Accordingly, failure to reject H_{01} by Stage 2 implies termination of the study and thereby corresponds to futility for H_{02} and H_{03}.

For $H_{02}: R > 1.3$, we assess the power of rejecting the null hypothesis when the true hazard ratio is under the alternative H_{A2} for $R = 1.15, 1.0, 0.9, \text{and } 0.8$.

At Stage 1 of 100 events, we assess H_{02} at $\alpha_{12} = 0.00005$.

At Stage 2, given no rejection of H_{02} at Stage 1, then at 200 events, we assess H_{02} at $\alpha_{22} = 0.00495$.

At Stage 3, given no rejection of H_{02} at Stage 1 or 2, then at 450 events, we assess H_{02} at $\alpha_{32} = 0.0100$.

At Stage 4, given no rejection of H_{02} at Stage 1, 2 or 3, but having conditional power by Stage 3 larger than 0.2 for Stage 4, then at 700 events, we assess H_{02} at $\alpha_{42} = 0.0175$ with a strategy like the Hochberg method with $\alpha_2 = 0.0225$. With this Hochberg method, we compare the maximum of p-values from Stage 3 and Stage 4 with α_2 as well as the p-value at Stage 4 with α_{42}; i.e., we reject H_{02} if either the maximum of the Stage 3 and Stage 4 p-values is less than α_2 or the Stage 4 p-value is less than α_{42}.

The cumulative power is computed for the rejection of H_{02} by Stages 1, 2, 3 and 4 respectively, with that by Stage 4 being the Type I error when H_{02} holds.

3.4 Analysis Plan for H_{03}: $R>1.0$

The rejection of H_{03} at any stage requires rejection of both H_{01} and H_{02} at that stage or a previous stage.

Accordingly, failure to reject H_{02} by Stage 4 implies termination of the study at Stage 4, and thereby corresponds to futility for H_{03}.

For H_{03}: $R > 1.0$, we assess the power of rejecting the null hypothesis when the true hazard ratio is specified under the alternative H_{A3} for $R_A = 0.9, 0.8, \text{and } 0.65$.

At Stage 1 of 100 events, we assess H_{03} at $\alpha_{13} = 0.00005$.

At Stage 2, given no rejection of H_{03} at Stage 1, then at 200 events, we assess H_{03} at $\alpha_{23} = 0.00005$.

At Stage 3, given no rejection of H_{03} at Stage 1 or 2, then at 450 events, we assess H_{03} at $\alpha_{33} = 0.00040$.

At Stage 4, given no rejection of H_{03} at Stage 1, 2 or 3, but conditional power by Stage 3 larger than 0.5 for Stage 5, then at 700 events, we assess H_{03} at $\alpha_{43} = 0.010$.

Table 3 Analysis Plan with Decision Rules

Stage	N_i	H_{01}:R>1.8	H_{02}:R>1.3	H_{03}:R>1.0
1	100	α_{11}=0.0125 → ↓	α_{12}=0.00005 → ↓	α_{13}=0.00005 ↓
2	200	α_{21}=0.0150 → α_1=0.025H	α_{22}=0.00495 → ↓	α_{23}=0.00005 ↓
3	450	N/A	α_{32}=0.0100 → CP≥0.2 ↓	α_{33}=0.00040 CP≥0.5 ↓
4	700	N/A	α_{42}=0.0175 → α_2=0.0225H	α_{43}=0.01000 CP≥0.5 ↓
5	900	N/A	N/A	α_{53}=0.02400 α_3=0.0245H

H= Hochberg for max p-value from last two stages (Hochberg, 1988).
CP= Conditional power for rejection at final stage. If insufficient, stop at current stage.
→=For each row, testing can proceed to the right after rejection of all hypotheses on the left at that stage or a previous stage.
↓= For each column, testing proceeds to the next row after no rejection of all hypotheses above.

At Stage 5, given no rejection of H_{03} at Stages 1, 2, 3 or 4, but having conditional power by Stage 4 larger than 0.5 for Stage 5, then at 900 events, we assess H_{03} at $\alpha_{53} = 0.024$ with a strategy like the Hochberg method with $\alpha_3 = 0.0245$. With this Hochberg method, we compare the maximum of p-values from Stage 4 and Stage 5 with α_3 as well as the p-value at Stage 5 with α_{53}; i.e., we reject H_{03} if either the maximum of the Stage 4 and 5 p-values is less than α_3 or the Stage 5 p-value is less than α_{53}.

The cumulative power is computed for rejection of H_{03} by Stages 1, 2, 3, 4 and 5 respectively, with that by Stage 5 being the Type I error when H_{03} holds.

The analysis plan for the five stage trial for all three hypotheses is illustrated in Table 3.

4. Simulation Plan

4.1 Simulation Plan to Generate Study Population

Simulations for the number of test treatment events at each stage are derived from binomial distributions, where the Binomial probability $\pi_A = \frac{R_A}{R_A+1}$ and R_A is the relative risk under the alterative hypothesis (or under the null hypothesis when computing the Type I error).

$y_i, i = 1, 2, 3, 4, 5$, is the number of events for the test treatment group among n_i total events, $n_i = 100, 100, 250, 250, 200$. y_i is generated from the $\text{Binomial}(n_i, \pi_A)$ distribution at the i^{th} stage. Let $g_i = \sum_{k=1}^{i} y_k$, $N_i = \sum_{k=1}^{i} n_k$, and $p_i = \frac{g_i}{N_i}$. Also let $\pi_{0j} = \frac{R_{0j}}{R_{0j}+1}$ for j=1, 2, 3 with $R_{0j} = 1.8, 1.3, 1.0$.

Table 4 Table for Generating Study Simulation Results

Stage	n_i	N_i	Distribution	Events	Total	Result
1	100	100	$B_1(100,\pi_A)$	y_1	g_1	Z_{1j} (q_{1j})
2	100	200	$B_2(100,\pi_A)$	y_2	g_2	Z_{2j} (q_{2j})
3	250	450	$B_3(250,\pi_A)$	y_3	g_3	Z_{3j} (q_{3j})
4	250	700	$B_4(250,\pi_A)$	y_4	g_4	Z_{4j} (q_{4j})
5	200	900	$B_5(200,\pi_A)$	y_5	g_5	Z_{5j} (q_{5j})

$\pi_A = \frac{R_A}{R_A+1}$, $R_A = 1.3, 1.2, 1.15, 1.10, 1.05, 1.0, 0.95, 0.9, 0.85, 0.8, \cdots$

$\pi_{0j} = \frac{R_{0j}}{R_{0j}+1}$, for $R_{0j} = 1.8, 1.3, 1.0$

$g_i = \sum_{k=1}^{i} y_k$, $N_i = \sum_{k=1}^{i} n_k$, $p_i = \frac{g_i}{N_i}$

$$Z_{ij} = \frac{[(p_i - \pi_{0j}) + 0.5/N_i]}{\sqrt{\pi_{0j}(1-\pi_{0j})/N_i}}$$

$q_{ij} = ProbNorm(Z_{ij})$

Compute the test statistics $Z_{ij} = \frac{[(p_i-\pi_{0j})+0.5/N_i]}{\sqrt{\pi_{0j}(1-\pi_{0j})/N_i}}$, and corresponding p-values $q_{ij} = ProbNorm(Z_{ij})$, where *i* is for the i^{th} stage and *j* is for the j^{th} hypothesis. Then, we compare q_{ij} to

the respective significance levels, α_{ij}, to decide whether to reject H_{0j}. The simulation plan for generating the study results is explained in Table 4.

4.2 Futility Rules

In a large trial investigating a therapy, especially in this rare events setting, interim analyses for early discontinuation due to futility might be recommended to reduce cost and potentially shorten the trial period. Conditional power is the probability of observing a statistically significant treatment effect at the end of a trial, conditional on the data observed thus far (Proschan & Hunsberger, 1995). As our study is a non-inferiority trial, we adapt the conditional power formula from the superiority hypothesis formula.

The conditional power is expressed as

$$\mathrm{f} = 1 - \mathrm{ProbNorm}\{(\sqrt{\frac{W}{1-W}} + \sqrt{\frac{1-W}{W}})\mathrm{Z} + \frac{Z_\alpha}{\sqrt{1-W}}\},$$

where *W* is the proportion of interim events to total events, *Z* is the observed test statistics at the interim analysis, *α* is the one-sided alpha level at the final analysis for each hypothesis (Lan & Wittes, 1988).

Conditional power for early discontinuation is calculated at Stage 3 for each of H_{02} and H_{03}, and also at Stage 4 for H_{03}, as are shown below.

At Stage 3, for a total of 450 events, we calculate

$$f_{32} = 1 - \mathrm{ProbNorm}\{(\sqrt{\frac{450/700}{250/700}} + \sqrt{\frac{250/700}{450/700}})Z_{32} + \frac{Z_{0.9825}}{\sqrt{250/700}}\}.$$

If $f_{32} \geq 0.2$, we continue to the 4th stage for testing H_{02}; otherwise, we stop the trial and conclude the hazard ratio is larger than1.3.

Also, we calculate $f_{33} = 1 - \mathrm{ProbNorm}\{(\sqrt{\frac{450/900}{450/900}} + \sqrt{\frac{450/900}{450/900}})Z_{33} + \frac{Z_{0.976}}{\sqrt{450/900}}\}$.

If $f_{33} \geq 0.5$, we continue to the 4th stage for testing H_{03}; else, we stop the trial for testing H_{03} at Stage 3 and conclude the hazard ratio is larger than1.0.

At Stage 4, for a total of 700 events, we calculate

$$f_{43} = 1 - \text{ProbNorm}\{(\sqrt{\frac{700/900}{200/900}} + \sqrt{\frac{200/900}{700/900}})Z_{43} + \frac{Z_{0.976}}{\sqrt{200/900}}\}.$$

If $f_{43} \geq 0.5$, we continue to the 5th stage for testing H_{03}; else, we stop the trial and conclude the hazard ratio is larger than 1.0.

5. Results

5.1 Simulation Results for Final Stage for All Three Hypotheses

Table 5 Simulation Results for the Final Stage for All Three Hypotheses

H_0	H_A	Power (No Futility Rule)	Power(Futility Rule)
>1.8	=1.8	0.023416	N/A
	=1.3	0.561306	N/A
	=1.15	0.840466	N/A
	=1.0	0.975183	N/A
	=0.9	0.9967	N/A
	=0.8	0.999798	N/A
>1.3	=1.3	0.022709	0.020167
	=1.15	0.319734	0.287184
	=1.0	0.901984	0.870683
	=0.9	0.994124	0.989791
	=0.8	0.999789	0.999725
>1.0	=1.0	0.023146	0.013049
	=0.8	0.908096	0.771679
	=0.65	0.999994	0.999043

For the strategies with and without the futility rules for the three hypotheses, the simulated Type I errors and power for the final analysis are shown in Table 5. If no futility rules are applied, the Type I errors for each hypothesis are 0.023, 0.023 and 0.023, respectively. With the futility rules, they are 0.023, 0.020, and 0.013. All Type I error rates are well controlled within the 0.025 significance level in either case. The power to reject each of the null hypotheses is satisfactory when their principal alternatives apply, as illustrated below.

5.2 Interim and Final Analyses for Each Hypothesis

For H_{01}: R>1.8, Stage 1 is the interim analysis, and Stage 2 is the final analysis. At Stage 1, we have power of 0.270 to reject the null hypothesis when the true hazard ratio is 1.3, and we have additional power of 0.291 from Stage 2 for a total power of 0.561 to reject H_{01}. Another hazard ratio of interest is 1.15; we have 0.500 power for early success at Stage 1, and additional power of 0.340 at Stage 2, for a total of 0.840 power for rejection of H_{01}. When the true hazard ratio is 1.0, 0.9 or 0.8, we have much better power to reject H_{01}, with any final power approaching 1.

Also, if we consider early success, H_{01} has power of 0.891 for rejection at Stage 1 when the hazard ratio is 0.9, and power of 0.986 when the hazard ratio is 0.8.

Table 6 Interim and Final Analyses for H_{01}: *R>1.8*

H_{A1}	Interim		Final
	1	2	
1.3	0.269814	0.291492	0.561306
1.15	0.499986	0.34048	0.840466
1.0	0.757683	0.2175	0.975183
0.9	0.890794	0.105906	0.9967
0.8	0.965527	0.034271	0.999798

For H_{02}: R>1.3, Stages 1, 2 and 3 are the interim analyses, and Stage 4 is the final analysis. When the true hypothesis is 1.15 and the futility rule is applied, we have 0.0003 power at Stage 1, 0.038 power at Stage 2, 0.118 power at Stage 3, and 0.131 power at Stage 4, for a total of 0.287 power for rejection of H_{02} by the final analysis. When the true hypothesis is 1.0 and the futility rule is applied, we have power of only 0.003 at Stage 1 to reject the null hypothesis, but we have an additional 0.215 power at Stage 2, power of 0.455 at Stage 3, and 0.198 power at Stage 4, for a total power of 0.871 power to reject H_{02} by the final analysis. When there is no futility rule applied at Stage 3, the power at Stage 4 increases from 0.0198 to 0.229 when R_A=1.0. We achieve better power when the true hazard ratio is 0.9 or 0.8.

Also, the power by Stage 3 is at least 0.99 when the true hazard ratio is 0.8, and it is nearly 0.8 by Stage 2.

Table 7 Interim and Final Analyses for H_{02}: $R>1.3$

H_{A2}	Interim				Final
	1	2	3	4	
With Futility Rules					
1.15	0.000309	0.038287	0.117679	0.130909	0.287184
1.0	0.003359	0.215058	0.454638	0.197628	0.870683
0.9	0.014232	0.473109	0.45297	0.04948	0.989791
0.8	0.054003	0.733579	0.209847	0.002296	0.999725
No Futility Rules					
1.15	*	*	*	0.163459	0.319734
1.0	*	*	*	0.228929	0.901984
0.9	*	*	*	0.053813	0.994124
0.8	*	*	*	0.00236	0.999789

*= No futility rule applies in the first three stages, and thus results are the same as the upper table.

For H_{03}: R>1.0, Stages 1, 2, 3 and 4 are the interim analyses, and Stage 5 is the final analysis. When the true hazard ratio is 0.8 and the futility rules are applied, we have 0.002 power at Stage1, 0.006 power at Stage 2, 0.132 power at Stage 3, 0.538 at Stage 4, and 0.094 power at Stage 5, for a total power of 0.772 by the final analysis to reject H_{03}. When the futility rules are not applied at Stages 3 or 4, the power at Stage 4 increases to 0.574 and the power at Stage 5 increases to 0.194, with the overall power by the final analysis increasing from 0.772 to 0.908. Much better power is achieved when the true hazard ratio is 0.65.

Table 8 Interim and Final Analyses for H_{03}: $R>1.0$

H_{A3}	Interim					Final
	1	2	3	4	5	
With Futility Rules						
0.8	0.002182	0.005543	0.132279	0.537821	0.093854	0.771679
0.65	0.032594	0.121436	0.707205	0.137479	0.000329	0.999043
No Futility Rules						
0.8	*	*	*	0.57395	0.194142	0.908096
0.65	*	*	*	0.138257	0.000502	0.999994

*= No futility rule applies in the first three stages, and thus results are the same as the upper table.

6. Conclusion

Our study design is well supported in that the Type I errors for all three of the null hypotheses are well controlled within the one-sided 0.025 level. Also, both the plans with and without the futility rules have sufficient power to reject the null if the true alternative is away from the null hypothesis as described previously in the tables of Section 5.

It is also noteworthy that the Type I errors without the futility rules are close to 0.025, which allows higher power under other alternative hypotheses.

Appendix

Table A Sample sizes using the Poisson and Logrank Methods

rr_0	*Method*	rr_a=1.0	rr_a =0.85	rr_a =0.80	rr_a =0.75	rr_a =0.70	rr_a =0.67	rr_a =0.60	rr_a =0.50
1	Poisson	N/A	1618.71	864.846	524.603	344.471	269.056	171.831	96.245
	Logrank	N/A	1598.296	851.101	514.864	337.405	263.317	168.119	94.567
	Poisson/Logrank	N/A	1.013	1.016	1.019	1.021	1.022	1.022	1.018
1.2	Poisson	2180.908	472.364	329.611	239.169	178.674	149.25	106.06	66.4566
	Logrank	2158.692	466.997	326.159	237.074	177.588	148.7	106.377	67.7076
	Poisson/Logrank	1.01	1.011	1.011	1.009	1.006	1.004	0.997	0.9815
1.3	Poisson	626.478	241.196	185.549	145.327	115.457	99.8233	75.215	50.3974
	Logrank	617.603	239.855	185.351	145.975	116.749	101.4606	77.4118	53.1938
	Poisson/Logrank	1.014	1.006	1.001	0.9956	0.989	0.9839	0.9716	0.9474
1.5	Poisson	266.213	136.4456	111.734	92.242	76.6511	68.0338	53.7022	38.0696
	Logrank	262.686	137.3426	113.437	94.5668	79.4624	71.1089	57.2071	42.0297
	Poisson/Logrank	1.013	0.9935	0.985	0.9754	0.9646	0.9568	0.9387	0.9058
1.8	Poisson	129.214	79.3731	68.0337	58.4819	50.3815	45.7004	37.5403	27.9969
	Logrank	128.716	81.76	71.0302	61.9724	54.2739	49.8167	42.0297	32.8901
	Poisson/Logrank	1.004	0.9708	0.9578	0.9437	0.9283	0.9174	0.8932	0.8512
2	Poisson	94.0439	61.6021	53.7332	46.9324	41.0275	37.552	31.3715	23.9208
	Logrank	94.5668	64.5342	57.2071	50.8559	45.3249	42.0612	36.2399	29.1873
	Poisson/Logrank	0.9945	0.9546	0.9393	0.9229	0.9052	0.8928	0.8657	0.8196

Note: rr_0 is the relative risk under the null hypothesis; rr_a is the relative risk under the alternative hypothesis; “Poisson/Logrank” indicates the ratio of the Poisson sample size to Logrank sample size.

Reference

Dann, R. S., Koch, G. G. (2004). Review and Evaluation of Methods for Computing Confidence Intervals for the Ratio of Two Proportions and Considerations for Non-inferiority Clinical Trials. *Journal of Biopharmaceutical Statistics, 15*(1), 85-107.

Denne, J. S., Koch, G. (2001). Monitoring a clinical trial with multiple hypotheses concerning the treatment effect on a single primary endpoint. *Statisics in Medicine, 20*(19), 2801–2812.

FDA. (2008). *Guidance for Industry: Diabetes Mellitus--Evaluating Cardiovascular Risk in New Antidiabetic Therapies to Treat Type 2 Diabetes.* Maryland.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportion* (2 ed.). New York: Wiley.

Freedman, L. (1982). Tables of the number of patients required in clinical trials using the logrank test. *Statistics in medicine, 1*(2), 121–129.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*(4), 800-802.

Journal: Lan, K. K., Wittes, J. (1988). The B-value: a tool for monitoring data. *Biometrics, 44*(2), 579-585.

Parks, M. H. (2009). *Clinical Perspectives on FDA Guidance for Industry:Diabetes Mellitus – Evaluating CV Risk in New Anti-diabetic Therapies to Treat T2DM.* Retrieved from www.fda.gov/downloads/Drugs/NewsEvents/UCM209087.pdf

Proschan, M., Hunsberger, S. (1995). Designed extension of studies based on conditional power. *Biometrics, 51*(4), 1315-1324.

Rothmann, M., Li, N., Chen, G., Chi, G. Y., Temple, R., Tsou, H.-H. (2003). Design and analysis of non-inferiority mortality trials in oncology. *Statistics in Medicine, 22*(2), 239-264.